Predicting Energy Performance Ratings of Domestic Properties in England and Wales: A Geostatistical Approach

36294908
MSc Data Science

A dissertation submitted for the degree of
Master of Science in Data Science

Supervised by *Professor, Christopher Jewell*

School of Computing and Communications
Lancaster University

September, 2023

Declaration

I declare that the work presented in this dissertation is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university.
Name: **36294908**
Date: **September, 2023**

Predicting Energy Performance Ratings of Domestic Properties in England and Wales: A Geostatistical Approach

36294908, MSc Data Science.
School of Computing and Communications, Lancaster University

A dissertation submitted for the degree of *Master of Science* in Data Science.

September, 2023

Abstract

Energy Performance Certificates (EPCs) play a crucial role in assessing building energy efficiency and guiding decisions for homeowners, tenants, investors, and policymakers. While prior research has made strides in predicting EPCs, it grapples with limitations, particularly the absence of a complete feature set necessary for evaluating energy efficiency in older homes lacking EPCs. This study offers a solution by introducing spatial modeling using Integrated Nested Laplace Approximation (INLA), which enhances prediction accuracy by considering aggregated features over modeling areas, addressing the challenge of missing data. Our results showcase a marginal improvement in predictions compared to the top-performing existing model (XGBoost), highlighting the effectiveness of spatial modeling in capturing intricate correlations. Bayesian inference, a pivotal aspect of our approach, provides comprehensive insights into energy efficiency, bolstering sustainable architecture and energy conservation initiatives, surpassing the capabilities of previous models' point predictions.

Acknowledgements

I extend my heartfelt appreciation to the individuals whose contributions were integral to the success of this dissertation project. Foremost, I express my gratitude to Professor Christopher Jewell, my academic supervisor, whose steadfast guidance and unwavering motivation were invaluable throughout this endeavor. I am also indebted to my host supervisor, John McHugh, for his unwavering support and invaluable insights. My sincere thanks go to Dr. Simon Tomlinson, my course engagement director, for affording me the opportunity to collaborate with the Centre for Energy Equality on this project.

I wish to convey my deep appreciation to my family for their unwavering prayers and support throughout this journey. Additionally, I extend my gratitude to my friends for their companionship during late-night report work and the enlightening discussions we shared.

I am profoundly thankful to all those who have been part of this endeavor, whether directly or indirectly, as your collective contributions have been instrumental in its accomplishment.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The motivation behind conducting research on the use of geospatial modelling for predicting energy performance efficiencies for domestic properties in the UK is rooted in the pivotal role of Energy Performance Certificates (hereafter referred to as EPCs). EPCs are not just official documents; they are a testament to a building's energy efficiency. Spawned from the Energy Performance of Buildings Directive (hereafter referred to as EPBD) in 2003, the birth of EPCs was the UK's response to mounting apprehensions over excessive energy consumption and its ensuing environmental ramifications. The EPBD's establishment, a directive from the European Union, was a strategic move to curtail energy consumption by buildings and elevate their intrinsic energy efficiency. Furthermore, it delineated guidelines for the architectural and constructional aspects of buildings, ensuring that the energy performance scores are not just visible but also easily interpretable.

EPCs cater to two primary needs. They grant clarity to potential homeowners, tenants, and investors about the energy performance of a building, thus empowering them with the knowledge to make judicious decisions concerning property acquisition or leasing. The ratings presented on these certificates are shaped by various factors, including the building's insulation standards, the efficiency of heating and air conditioning systems, and the structure's age. Additionally, EPCs underpin the UK government's policy-making apparatus, particularly policies promoting energy conservation and carbon footprint reduction. The transparency and measurability of energy performance offered by EPCs stimulate refurbishments and alterations to pre-existing structures, steering the nation towards a future characterized by sustainable architecture.

Yet, the impetus for this research goes beyond the foundational role of EPCs. The UK's residential sector has seen a significant fraction of its properties being subjected to energy efficiency evaluations. Data reveals that over half of dwellings in England and in Wales have had an EPC since the initiation of these records (see Figure 1.1). Despite these numbers, it's

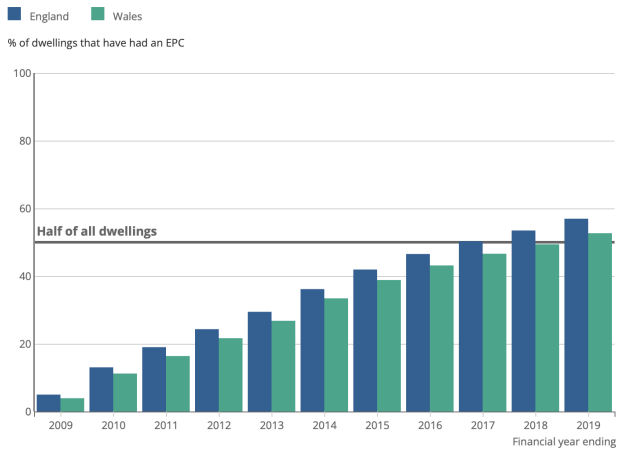

Figure 1.1: Estimated proportion of the dwelling stock that has had at least one Energy Performance Certificate, financial year ending 2009 to financial year ending 2019, England and Wales

disconcerting to note the absence of energy efficiency assessments for a sizable fraction of older homes. This gap in EPC coverage for such properties impedes a comprehensive evaluation of their energy efficiency and poses challenges in pinpointing avenues for enhancement.

Researchers from the Data Science Campus, have made significant strides in addressing this problem. Their investigations delved into the EPC data for Wales and the creation of diverse machine learning models tailored to the data (Williams and Bonham, 2020). However, their work did not specifically probe the potential spatial correlation of energy efficiencies across the UK. By leveraging their foundational research and integrating geospatial modelling, this study strives to enhance accuracy and provide deeper insights into predicting energy performance ratings.

1.2 Project Aims

The primary focus of this research project revolves around several pivotal objectives. Firstly, it seeks to delve into the findings of preceding studies, aiming to discern any evidence of spatial correlation in the distribution of energy efficiencies across the entirety of the UK. Such an exploration will provide insights into regional disparities or patterns in energy efficiency metrics. Secondly, based on the insights gathered, the project endeavors to pinpoint and subsequently implement a spatial modelling approach best suited for this dataset and context. This methodology will be instrumental in capturing the nuances of energy efficiency distribution geographically. Lastly, a significant portion of this research will be dedicated to a comparative analysis. The newly developed spatial model's efficacy will be compared against the models adopted in prior research. Through this comparison, the research will

shed light on the advancements made and the depth of understanding achieved regarding energy performance ratings prediction.

1.3 Overview of the Report

The report is organized into distinct chapters, each contributing to the understanding of geospatial modeling for predicting energy performance efficiencies:

- Literature Review: This chapter introduces energy efficiency modeling, reviews prior research on Energy Performance Certificates (EPCs), and emphasizes the relevance of geospatial modeling in energy efficiency studies.
- Methodology: This chapter details data collection procedures, explains geospatial modeling techniques, outlines predictive model development, introduces Gaussian Spatial Processes, and describes model fitting and prediction steps.
- Results: In this chapter, research outcomes are presented, including predictions, evaluation metrics, feature importance analysis, model comparisons, prediction visualizations, and an evaluation of prediction accuracy and errors.
- Discussion: This chapter interprets feature effects on energy efficiency predictions, conducts comprehensive model performance comparisons and evaluations, and reflects on overall project findings and their implications.
- Limitations and Future Works: Addressing research limitations and proposing potential future refinements and improvements in energy efficiency modeling are the main focus of this chapter.
- Conclusion: This concluding chapter summarizes key findings, underscores the significance of geospatial modeling in energy efficiency predictions, and highlights practical implications derived from the research.

Chapter 2

Literature Review

2.1 Related Work

EPCs were introduced to assess the energy efficiency of properties. These certificates play a pivotal role in assisting governments in understanding their current energy landscapes, formulating policies to achieve carbon reduction goals, and recommending energy efficiency enhancements. In 2008, the UK government made it mandatory to possess EPCs for properties, thereby establishing them as compulsory prerequisites for construction, sale, or rental (EPC Regulations, 2007). However, this regulation resulted in a substantial absence of EPCs for properties constructed before 2008. This data gap poses significant constraints on effective policy-making related to emissions and energy matters.

The Energy Performance of Buildings Register (hereafter referred to as EPBR) contains comprehensive data on issued EPCs, encompassing property ratings and factors contributing to the energy efficiency score. This dataset serves as a valuable resource for training machine learning models to predict Energy Performance Ratings (hereafter referred to as EPRs) for properties lacking EPCs. However, one critical prerequisite for such predictions is having complete feature values used in the modeling process. This feature set consists of both numerical and categorical attributes. Numerical attributes primarily pertain to physical property dimensions, including energy consumption, floor area, floor height, and the number of rooms. Conversely, categorical features encompass details regarding construction materials and heating systems, such as roof type, floor type, and the type of main heating system.

Addressing this data gap, previous research explored the application of machine learning techniques, including tree-based algorithms and neural networks, for modeling the EPC dataset (Williams and Bonham, 2020). Their work also involved identifying potential proxies capable of accurately substituting hard-to-obtain feature values from the actual EPC feature set. While values for structural property dimensions and proxies for derived features can be sourced from Ordnance Survey datasets (OS Data Hub), acquiring proxies for features related to building materials and construction types presents challenges.

Recognising that houses within the same neighborhoods often share similar construction attributes, their research proposed the use of geographic features as suitable proxies. Geographic attributes, such as latitude, longitude, and the energy efficiencies of neighboring households, were suggested as proxies for these qualitative features. Experimental results with the proxy feature set identified the XGBoost regressor as the most effective model.

2.2 Spatial Modelling

It is noteworthy that the model trained on the proxy feature set did not exhibit the same level of performance as the one trained on the actual feature set. Additionally, their approach involved the inclusion of geographic attributes in the proxy feature set to account for missing qualitative property features and enable the learning of spatial relationships among neighboring properties (Williams and Bonham, 2020).

Instead of depending on non-spatial models to incorporate spatial aspects, a more effective strategy involves the utilisation of explicit spatial models for prediction (Amara and Ayadi, 2012). This strategic shift provides a remedy for the challenge of identifying suitable proxy features. By harnessing a spatial model, predictions can be generated for small, localised regions encompassing properties with missing data. The feature set aggregated over these regions acts as an approximate representation of the typical property within each region. Significantly, as the size of these regions diminishes, prediction accuracy improves significantly (Christiaensen et al., 2011). This approach not only obviates the need for proxies but also capitalises on the inherent spatial dependencies in the data, resulting in more precise predictions for properties with incomplete information.

Chapter 3

Methodology

This chapter offers a detailed overview of the analytical methods and procedures employed in this investigation. This study presents a comprehensive exploration of the techniques employed to achieve the research objectives, encompassing various facets such as data collection, preprocessing, analysis, and modelling. This section not only offers a clear rationale for the chosen methodology but also lays the foundation for a systematic and meticulous study. This chapter provides an exhaustive description of the processes used to investigate geographical patterns, perform variance analysis, and apply predictive modelling techniques. It ensures transparency by delineating the systematic and replicable approach employed to uncover meaningful insights.

3.1 Data

Within this section, the foundational aspects of the study are explored by engaging in a comprehensive discussion of the origins of both the primary and supplementary datasets. This involves explaining essential concepts inherent in the dataset. This section establishes the groundwork for a profound understanding of the data that underpins our research efforts, explaining the data's origins and introducing key terminology.

3.1.1 Energy Performance Certificates

EPCs provide comprehensive information regarding energy usage and estimated energy costs associated with properties. Additionally, they offer recommendations for reducing energy consumption and achieving financial savings.

The possession of EPCs is a mandatory requirement for newly constructed, sold, or rented houses. Before initiating the marketing process for the sale or rental of a property, property owners must obtain an EPC to provide to prospective buyers or tenants.

An authorised evaluator assesses the property and subsequently issues an EPC, indicating the efficiency rating on a scale ranging from A (representing high efficiency) to G (indicating lower efficiency), as shown in Figure 3.1. The EPC remains valid for 10 years from the date of issuance.

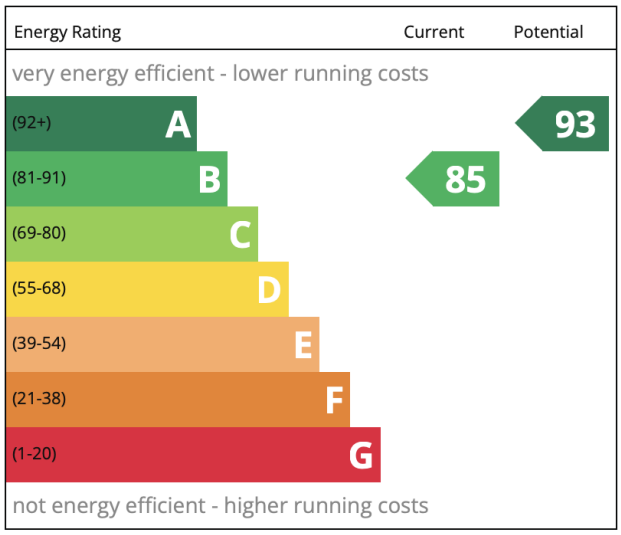

Figure 3.1: Energy Efficiency Rating for a property (taken from Energy Performance of Buildings Register)

3.1.2 Energy Performance of Buildings Register

The Energy Performance of Buildings Register managed by the Department for Levelling Up, Housing and Communities (DLUHC), encompasses EPCs for both residential and non-residential properties situated in England and Wales. This dataset is made accessible under an open licence and is available to those who register on the department's website.

A user-friendly tool, presented in the form of an interactive dashboard, is provided to users, allowing them to filter and download EPC certificates. The primary dataset utilised for this project is derived from the aforementioned source.

3.1.3 Output Areas

Output Areas (OAs) were introduced following the 2001 Census as the most detailed geographical units used for census data purposes (National Statistics, 2021). Generally, an OA is characterized by a cluster of approximately 40 to 250 houses, accommodating a resident population ranging from 100 to 625 individuals. Lower layer Super Output Areas (hereafter referred to as LSOAs) are often comprised of clusters of OAs, which commonly

consist of four or five OAs. These LSOAs comprise an estimated range of 400 to 1,200 houses, accommodating a resident population of approximately 1,000 to 3,000 individuals. 33,755 LSOAs are located in England, and 1,917 are located in Wales.

The Open Geography Portal (*Open Geography portalx* n.d.), which is provided by the Office for National Statistics (ONS), grants users unrestricted and complimentary access to a range of geographic resources, including census borders, maps, and look-up tools. It is possible to filter and download the LSOA boundaries that correspond to the region of interest from *Lower Layer Super Output Areas (2021) Boundaries EW BFC* 2023. Additionally, a dataset containing the mapping of postcodes to LSOAs is available for download from the postcodes lookups section (pcd_oa_lsoa_msoa_lad lookup).

3.1.4 Ordinance Survey Data Hub

The United Kingdom's national mapping agency, Ordnance Survey (OS), holds the key responsibility of creating, maintaining, and providing geographic information, maps, and data. The OS Data Hub (OS Data Hub) is an online platform offered by OS, granting users access to a diverse range of geospatial data and mapping services. Within this hub, the OS National Geographic Database (NGD) stands as a comprehensive repository of geospatial data, encompassing various geographical features and properties across the United Kingdom. The OS NGD application programming interfaces (APIs) provide a convenient means for querying and accessing the NGD.

For this project, two NGD APIs can be utilised. These APIs enable the retrieval and incorporation of necessary data that may be missing from some of the EPC records. The Features API can be employed to gather information about the structural dimensions of households and their geographic boundaries. The Linked Identifiers API facilitates the establishment of connections among different identifiers used in the mapping process. In this specific use case, the Linked Identifiers API can be applied to establish a correspondence between the Unique Property Reference Number (UPRN) of a property and the Topographic Identifier (TOID) associated with the features. The combination of these two APIs presents an effective solution for retrieving any missing information within the EPC records.

3.1.5 Dataset

The Energy Performance of Buildings Register contains a total of 24,894,036 EPCs for residential structures located in England and Wales. Upon download, the data is meticulously organised into 345 folders, each corresponding to specific local authorities.

Each individual EPC record encompasses the following key attributes:

1. **Current and Potential Energy Efficiency Scores:** These scores, expressed as integers ranging from 0 to 100, provide a numerical representation of the annual energy expenditure associated with the property. The potential energy efficiency score offers

an estimate of the property's potential performance after implementing recommended enhancements.

2. **Current and Potential Energy Ratings:** Utilising a linear scale denoted by letters from A to G, these ratings evaluate the energy efficiency of the property. The 'A' rating signifies the highest level of achievable energy efficiency.

3. **Geographical Location Details:** Inclusive of particulars such as address and postcode, these details offer insights into the property's spatial context.

4. **Structural Characteristics:** Encompassing metrics like floor area and floor height, this category furnishes information about the physical dimensions of the property.

5. **Construction Material Specifications:** This section provides a detailed description of the materials used in constructing various components, including windows, floors, walls, and the roof.

6. **Energy Consumption Patterns and Heating Infrastructure:** Comprising data on energy consumption, heating mechanisms, and control systems, this segment offers a comprehensive overview of the property's energy-related attributes.

More detailed explanation about the features are given in the appendix.

3.2 Data Preprocessing

In this section, a crucial endeavour is undertaken to refine and augment the unprocessed dataset. This area contains a variety of essential operations, including as data sampling, dimensionality reduction, cleaning procedures, categorical feature binning, transformation strategies, and the practise of integration. Every individual stage in the data preparation process is crucial for conducting a comprehensive analysis. These steps aim to minimise the impact of noise, efficiently handle categorical information, and turn the data into a format that facilitates the extraction of important insights. By thoroughly examining these crucial subjects, this section establishes the groundwork for the succeeding stages of analysis, facilitating a full and perceptive investigation of the dataset.

The preprocessing steps for removing bilingual and Welsh sentences and categorical feature binning using CHAID have been reused for this study from (Williams and Bonham, 2020).

3.2.1 Sampling

Given the vast scope of the EPC dataset obtained from the EPC register, covering homes across England and Wales, a prudent approach to data sampling was employed. The objective

was to streamline the analysis by focusing on a specific subregion within the United Kingdom, thereby enabling more precise forecasts.

In this project, a well-defined subregion was delineated, as illustrated in Figure 3.2, encompassing dynamic urban centres such as Manchester, Liverpool, Leeds, Sheffield, and Lancaster. The subregion was carefully selected using the polygon selection tool provided by the ONS open geography webpage. To establish the spatial scope, the borders of LSOAs within this subregion were utilised.

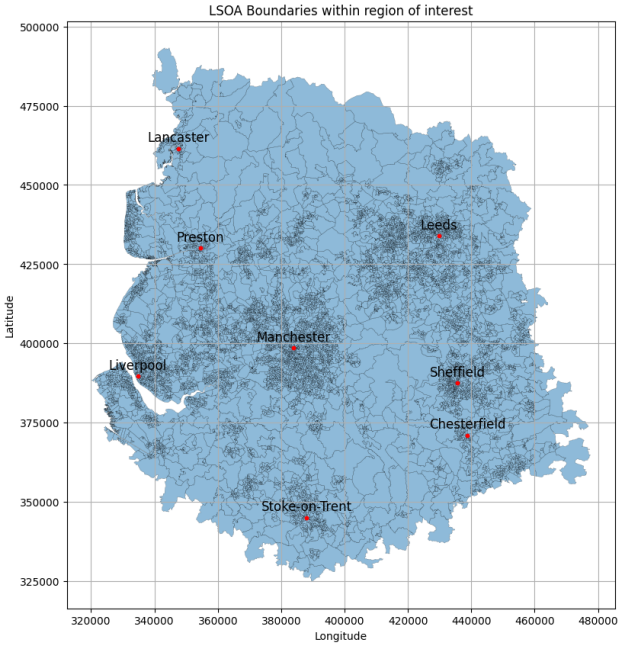

Figure 3.2: Sampled region

By opting to operate within this specific subregion, the dataset's extent was significantly narrowed, while still encompassing crucial urban areas. To construct a comprehensive dataset suitable for study, EPC records corresponding to all local authorities falling within the boundaries of the LSOAs were extracted. It is important to note that certain properties possess multiple EPCs over a given period. To ensure the dataset represents the most recent property assessments, the latest EPC was selected for each property in such cases.

This focused data sampling approach lays the foundation for the subsequent analysis and modelling work. It also enables predictions regarding energy performance efficiency within a well-defined and meticulously chosen subregion. The distribution of energy efficiency among properties in the sampled subregion is depicted in Figure 3.3.

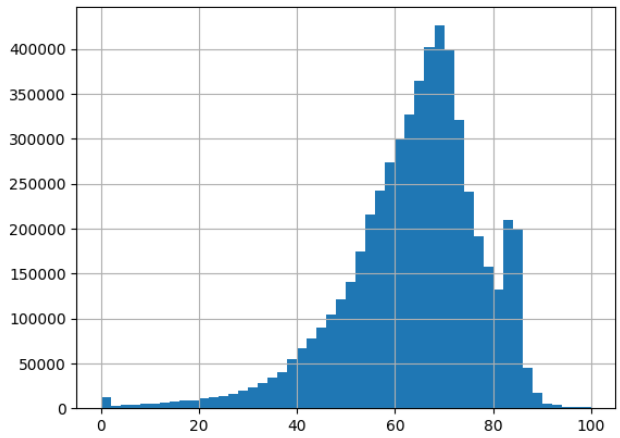

Figure 3.3: Distribution of energy efficiencies of houses in sampled subregion

3.2.2 Dimensionality Reduction

Dimensionality reduction was a crucial aspect of data preprocessing in the EPC dataset. The dataset exhibited a diverse range of properties, comprising a thorough collection of 92 elements. Nevertheless, it should be noted that not all parameters possessed inherent relevance when it came to modelling EPRs. A pragmatic methodology was employed to ascertain and preserve characteristics that possess true potential to contribute to useful predictions.

Certain properties demonstrate the potential values associated with features such as `CO2_EMISSIONS_POTENTIAL` and `LIGHTING_COST_POTENTIAL`. These values lack statistical significance in relation to EPR predictions. Likewise, attributes such as `HOT_WATER_ENV_EFF`, `WINDOWS_ENV_EFF`, `WALLS_ENV_EFF`, and other indicators representing environmental efficiency were identified, although they did not possess direct relevance to the EPR modelling process.

Furthermore, several qualities were deemed challenging to assess, namely `HEAT_LOSS_CORRIDOR` and `UNHEATED_CORRIDOR_LENGTH`. In order to prioritise practicality and market applicability, it was determined that the model would be built using readily available and substantively significant characteristics.

In order to enhance the efficiency of the dataset, features that exhibited null values exceeding 50% were excluded. As a result, several features were omitted from the analysis, including `LODGEMENT_DATE`, `LODGEMENT_DATETIME`, `INSPECTION_DATE`, `POTENTIAL_ENERGY_RATING`, and various others specified in the original dataset.

Through a meticulous method of refining the feature set, the dataset was carefully curated to contain features that hold significant importance, hence aligning with the practical application of the prediction model. The process of dimensionality reduction employed in this exercise enhances the suitability of our dataset for making EPR forecasts that are both precise and significant.

3.2.3 Cleaning

Data cleaning stands as a pivotal process in data management, dedicated to enhancing the quality and consistency of data.

Despite being assessed by trained assessors, the EPC dataset exhibited outliers that had the potential to impact the accuracy of predictions. To enhance the precision of the predictions, it became imperative to execute a critical data cleaning phase that duly accounted for the influence of outliers.

Clipping Numeric Features: To address the influence of outliers, the numeric features underwent a clipping procedure. This involved setting an upper bound at the 95^{th} quantile and a lower barrier at zero (or the minimum value for the feature). This approach effectively mitigated the impact of outliers. It's worth noting that due to the intrinsic non-negativity of all characteristics, assuming a minimum value of zero was considered a safe and appropriate choice.

Addressing Correlation: The identification of highly linked traits has emerged as a crucial step. The existence of associated qualities might lead to the inclusion of unnecessary dimensions in a model, without providing significant modelling advantages. In order to examine this issue, the Pearson's correlation coefficient was computed for the numerical variables, revealing pairs of variables that are correlated (see Figure 3.4). Scatter plots were utilised in order to conduct a more in-depth examination of the linear associations among features that exhibited a correlation coefficient beyond 0.6, regardless of whether the connection was positive or negative. Subsequently, these associations were further investigated, shedding light on the nature of the correlation.

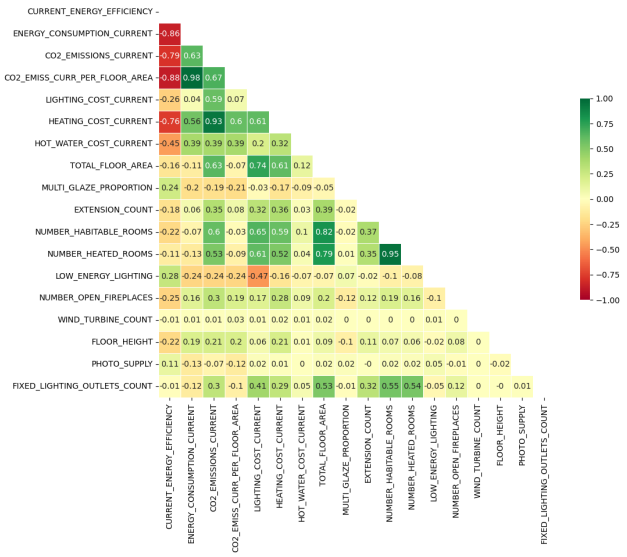

Figure 3.4: Correlation between numeric features

To enrich this analysis, chi-squared tests for independence were conducted to explore the interplay among categorical variables. This complementary approach enhanced the investigation, elucidating intricate correlations and fostering a more nuanced understanding of the dataset's interdependencies.

Managing Categorical Features: The dataset comprised categorical features, with a prevalence of Welsh words observed in homes located in close proximity to Wales. To ensure uniformity, Welsh nouns were replaced with their English equivalents, and multilingual sentences underwent the removal of Welsh segments. Additionally, the values related to "Average thermal transmittance" were standardized to ensure that the units were adjusted to W/m^2K, maintaining consistency throughout.

Mitigating Missing Values: The mitigation of missing values was crucial in addressing the absence of numeric feature values. The process of imputation involved substituting missing data with the average value of the corresponding feature for properties located within the identical postcode region. This methodology ensured the preservation of data integrity while also improving the accuracy of prediction outcomes.

By implementing rigorous data cleaning techniques, the integrity of the dataset was enhanced, hence coinciding with the goals of exact modelling and accurate projections of Energy Performance Ratings.

3.2.4 Categorical Feature Binning

The EPC dataset comprises numerous categorical variables, several of which exhibit a wide range of distinct values. Textual descriptions, often created individually, aim to represent the diversity of properties. However, these descriptions may lead to the assignment of unique values to a single property. The inclusion of categorical variables with a high number of dimensions poses challenges in the modeling process, resulting in heightened complexity and reduced interpretability.

Model Complexity and Interpretation: One of the primary challenges arises from the growing complexity of models and the diminishing interpretability when dealing with a substantial number of levels in categorical variables. This situation necessitates the incorporation of additional parameters to accurately capture the underlying relationships, potentially leading to overfitting and impeding generalisation. The intricate nature of interpreting the influence of various factors on the dependent variable poses a challenge, constraining the practical insights that can be derived from the model.

Binning Strategy: The binning strategy, also known as discretisation or grouping, offers a strategic solution to tackle the complexities presented by these challenges. By reducing the number of levels and consolidating them, the binning process simplifies variables, thereby reducing the model's complexity and enhancing its overall performance. The strategic alignment approach helps mitigate the risk of overfitting and facilitates the achievement of clear and comprehensible results. Additionally, binning addresses the challenges posed by

high-dimensional data, improving the capacity to handle and analyse large datasets while enhancing the accuracy of predictions on new data.

Utilizing the CHAID Algorithm: The Chi squared Automatic Interaction Detector (hereafter referred to as CHAID) algorithm is employed in this study for categorical feature binning. The CHAID technique, widely recognized and employed in decision tree analysis, automatically divides categorical variables into distinct groups with comparable response rates (Kass, 1980). This segmentation process is specifically tailored to align with the EPRs. It enhances the similarity of values within a group for the variable of interest while optimising differences between different groups.

Moreover, the utilisation of CHAID's automated methodology proves particularly advantageous given the limited availability of comprehensive domain expertise. Manual binning carries inherent risks associated with potential improper grouping, errors, and biases. CHAID's objectivity stems from its reliance on statistical significance and objective criteria, surpassing subjective judgments.

Binning Process: The binning procedure involves dividing EPR into two distinct categories: the first quantile (0-25th quantile) and fourth quantile (75-100th quantile). This approach excludes the intermediate quantiles (25-75%) to facilitate the differentiation of the target variable. A CHAID tree is computed at a single level for each categorical feature, creating CHAID nodes to represent the resulting groupings. The exclusion of some levels of categorical variables corresponding to middle quantiles is primarily due to their infrequency. The assignment of these values to one of the groups generated by CHAID is conducted with careful consideration, guided by professional judgment.

The existing categories are documented in a dictionary-style format, subsequently used to replace individual levels with their respective groups. Through a systematic methodology, the process of categorising features into bins enhances the dataset's robustness, establishing a more sophisticated foundation for predictive modeling and generating informative forecasts for Energy Performance Ratings.

3.2.5 Transformation

All numeric features, except the target variable `(CURRENT_ENERGY_EFFICIENCY)`, undergo a min-max scaling transformation to be within the range of 0 to 1. Min-max scaling is employed in this context as it preserves the relative relationships between data points and prevents the dominance of larger-scale features in the model. Given that all numeric features have been clipped to limits, min-max scaling can be applied without concerns about the impacts of outliers.

3.2.6 Integration

The primary dataset lacks any geographic attributes, such as property boundaries. In the context of spatial modeling, the inclusion of geographic borders or coordinates is an essential requirement.

While utilising property boundaries would have been a logical approach, considering each data point in the dataset corresponds to a distinct household, constraints emerged. Despite the potential use of the OS NGD features API and linked identifiers API to establish a correlation between each property and its respective boundary via the UPRN, the imposed rate limit on the API posed a time constraint in accessing this data for the multitude of homes within the specified territory. Furthermore, even in the scenario where boundaries for every residence are obtained, the process of fitting a model to such a vast dataset, comprising over 3 million entries, would entail substantial computing costs.

A recommended course of action would have involved consolidating the data related to postcodes and utilising it for modeling purposes. The boundaries of postcodes can be acquired through the utilisation of the OS Code-Point with Polygons product. Unfortunately, access to this specific product is restricted solely to OS licensed partners and institutions covered by the Public Sector Geospatial Agreement (PSGA).

Consequently, the decision was made to utilise the LSOA borders dataset. LSOAs are explicitly structured to maintain relatively uniform population sizes, rendering them suitable for statistical examination and comparative analyses. They more accurately depict local neighborhoods and contribute to preserving individual privacy by aggregating data at a higher level. Implementing this measure effectively mitigates the potential release of confidential data concerning specific individuals while simultaneously ensuring the availability of valuable insights at a regional scale.

As a result, it was decided to group the dataset at the LSOA level and employ it for geographic modeling. Thus, the EPC dataset is aggregated based on LSOA regions. All numeric features are assigned the mean value corresponding to the LSOA, and categorical features take on the most frequently observed category in that LSOA.

However, the challenge of associating each attribute with its respective LSOA remains unresolved. The provided dataset contains postcodes corresponding to each property. The integration of LSOA boundaries into the real dataset is facilitated by utilising the postcode LSOA lookup dataset.

3.3 Spatial Variance Analysis

To demonstrate the necessity of spatial modeling in this project, a thorough analysis of spatial variance was undertaken. This analysis plays a crucial role in comprehending the spatial dependencies present within the dataset. The subsequent series of actions illustrates the approach:

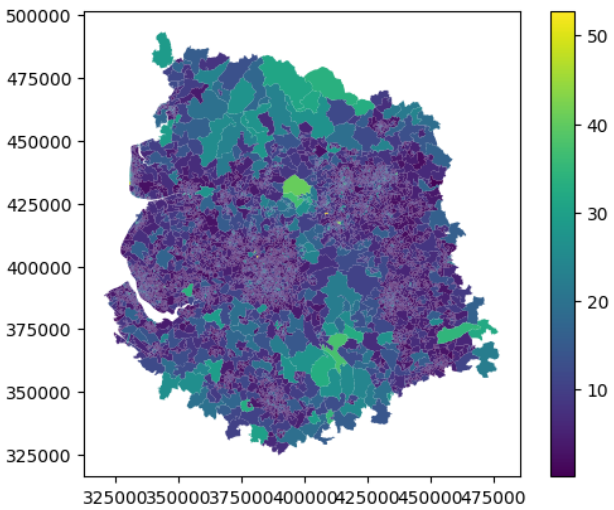

Figure 3.5: Spatial distribution of Loss

3.3.1 Loss Calculation from XGBoost Model

In the previous work (Williams and Bonham, 2020), it was determined that the XGBoost regressor model exhibited the highest prediction accuracy for this dataset. Consequently, in the initial phase, the XGBoost model underwent training using households within the specified region, and subsequently, the mean squared loss was computed. This step offered fundamental insights into the predictive accuracy of the model across the geographic landscape.

3.3.2 Visualization of Loss and Spatial Correlation

The computation of mean loss values entailed aggregating prediction loss across the LSOAs. To visually represent the distribution of mean squared loss values across the designated geographic region, a geographic mapping technique was employed. The primary goal of this visualisation was to uncover spatial patterns by depicting regions with both high and low loss values. The study aimed to investigate whether discernible spatial correlation patterns could be identified through visual analysis.

The study unveiled the presence of localised areas exhibiting low loss values in proximity to urban centers. These areas gradually transitioned into regions characterised by higher loss values as one moved towards the periphery (refer to Figure 3.5).

3.3.3 Variogram

The variogram serves as a vital statistical tool within the realm of spatial analysis and geostatistics, offering a means to quantify the spatial variability or dependence of a variable across a designated geographic area. This metric assesses the extent to which data values' similarity either increases or decreases concerning the spatial separation between data collection points.

The variogram graph visually represents the average variance in values among pairs of points relative to their spatial separation, often referred to as lag distance. The lag distance denotes the spatial gap between two points where the variance calculation occurs.

To infer spatial correlation from the variogram, the following patterns are observed:

Absence of Spatial Correlation: When the variogram tends to stabilize and reach a plateau as lag distance increases, it indicates the absence of spatial correlation. This pattern suggests that variations in values among points do not depend on their spatial proximity.

Spatial correlation: A positive relationship between the variogram and lag distance signifies an increase in spatial dependence. This suggests that points further apart exhibit greater disparities in their values compared to points in close proximity. Observing the variogram's behavior can help determine the scale of spatial correlation, indicating the distance required between points for them to exhibit no correlation.

3.3.4 Statistical Analysis Using Variograms

To further substantiate the findings, a rigorous statistical analysis was conducted employing variograms (Carr, Bailey, and Deng, 1985).

A dataset for variogram calculations was generated by combining the LSOA centroids and the aggregated mean loss over these LSOAs. A permutation test was employed to assess the statistical significance of the observed pattern compared to patterns under the null hypothesis.

In this test, the null hypothesis postulates that the dataset exhibits no spatial variance, while the alternative hypothesis suggests the presence of spatial variance within the data.

The null hypothesis encompasses 1000 samples from the variogram dataset. Each sample is constructed by reshuffling the loss values while keeping the centroid coordinates constant. Variance is calculated for each of these samples (refer to Figure 3.6). The 2.5% and 97.5% quantiles for the variances across all 1000 samples are computed. The red lines represent the quantiles at various spatial lags.

Subsequently, the variogram is plotted for the actual unshuffled dataset (see Figure 3.7), serving as the sample for the alternative hypothesis. The quantiles derived from the null hypothesis are superimposed onto this variogram.

The variogram analysis yielded valuable insights. The variance plot exhibited distinctive patterns:

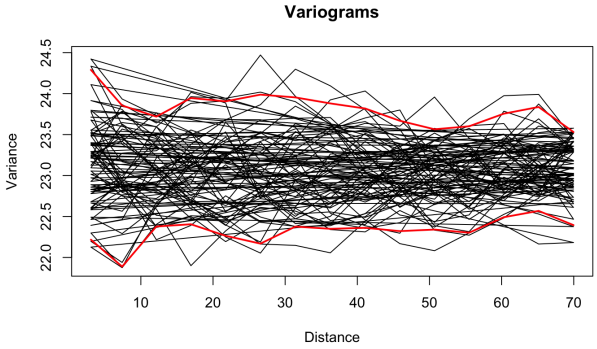

Figure 3.6: Variograms from the null hypothesis

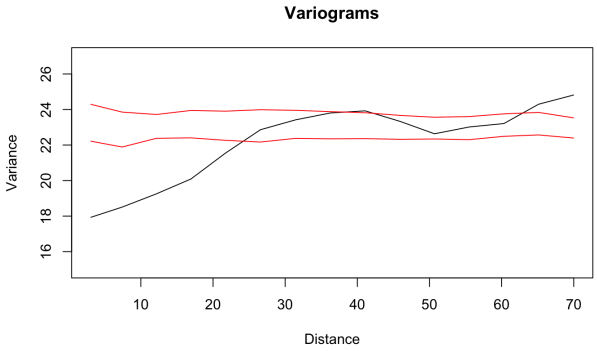

Figure 3.7: Actual variogram overlayed with quantiles from null hypothesis samples

- An initial phase with small variance values that increased up to the 2.5% quantile.
- Within the 2.5% to 97.5% quantile range, the variance remained relatively stable, signifying a consistent spatial correlation zone.
- Beyond the 97.5% quantile, the variance exhibited another notable increase, suggesting a stronger spatial correlation at larger distances.

These observed patterns and trends in the variance analysis offer compelling evidence of spatial dependence within the dataset. The variance in loss values at varying distances underscores the presence of spatial relationships. Urban centres, with their high population density and similar EPRs, exhibit minimal variance in loss. As we approach city borders, population density and EPRs decline. Moving towards the outskirts, fewer properties share nearly identical EPRs. This explains the variogram's progression from low variance to the 2.5% quantile, followed by a plateau till 97.5%. Continuing outward, the analysis encounters the boundaries of other cities, marked by differing average EPRs and a growing EPR gradient towards city centres. These factors contribute to the variance increase beyond the 97.5% quantile.

3.4 Spatial modelling using INLA

3.4.1 Type of data

In the realm of spatial data, various types exhibit unique attributes (Cassie, 1993):

Areal data is linked to predefined geographic regions, like counties or administrative zones, and primarily deals with aggregated attributes, such as counts or rates.
Geostatistical data originates from continuous observations associated with specific geographic coordinates, commonly used for interpolation and continuous surface models.
Point patterns involve representing individual occurrences or objects as points, facilitating the analysis of spatial randomness or clustering.

In this context, each household represents a distinct area or region within the United Kingdom, with continuous energy efficiency values linked to these areas. Associating each value with a specific area classifies the data as areal. Due to the irregular boundaries of these areas, the data falls into the category of irregular lattice areal data.

3.4.2 Gaussian Processes (GPs)

Gaussian Processes (hereafter referred to as GPs) provide an attractive framework for modeling and predicting irregular lattice areal data. They excel in estimating uncertainty, capturing spatial dependence, handling outliers, and adapting to irregular data patterns

(Banerjee, Dunson, and Tokdar, 2012). GPs are adept at estimating uncertainties in real-world scenarios with inherent variability, measurement errors, and unobserved factors. They offer both point predictions and confidence intervals, which are crucial for understanding prediction reliability (Banerjee, Dunson, and Tokdar, 2012). GPs accurately account for spatial autocorrelation by emphasizing the similarity between nearby data points, making them well-suited for datasets with irregular spatial structures. They also exhibit robustness in the presence of outliers, ensuring reliable performance even with unexpected observations.

One notable advantage of GPs is their ability to extrapolate into unobserved regions and interpolate between observed points, making them valuable for predicting in areas with limited data. Additionally, GPs have a Bayesian nature, enabling seamless integration of prior knowledge and continuous updating of predictions as new data becomes available.

Furthermore, GPs exhibit spatial continuity, allowing them to effectively model each individual property within the extensive dataset. However, it's essential to consider the computational challenges associated with large datasets. The dimensions of the spatial covariance matrix can become substantial due to the numerous properties in the dataset. Computing matrix inversions, which have a time complexity of $O(N^3)$, can be particularly demanding with such massive matrices. Therefore, strategic computational approaches are necessary to ensure practical implementation.

3.4.3 INLA

The Integrated Nested Laplace Approximation (hereafter referred to as INLA) proves to be an excellent choice for implementing GP on irregular lattice areal data. It leverages integrated nested Laplace approximation (Rue, Martino, and Chopin, 2009), a method renowned for its precision in approximating complex posterior distributions. In terms of computational complexity, this method is more efficient compared to traditional Markov Chain Monte Carlo (MCMC) techniques (Rue, Martino, and Chopin, 2009). While MCMC relies on iterative sampling, INLA swiftly computes posterior distributions using a nested sequence of Laplace approximations. This computational speed is particularly advantageous when working with large spatial datasets.

All these features of INLA is available through the R-INLA package (Rue, Lindgren, et al., n.d.). Moreover, the Bayesian approach used in R-INLA offers valuable uncertainty estimates, which are essential for making informed decisions in spatial contexts where uncertainties may stem from sources like measurement errors or unobserved spatial factors. In summary, R-INLA provides a rapid, precise, and user-friendly platform for applying Gaussian Processes to irregular lattice areal data, facilitating comprehensive spatial modeling and prediction.

3.4.4 Model Explanation

R-INLA operates within a Bayesian paradigm (Lindgren and Rue, 2015), offering a robust approach to spatial modeling of areal data. The Besag, York, and Mollié (hereafter referred to as BYM) model within INLA consists of a convolution between an intrinsic Conditional Autoregressive (CAR) model and an independent and identically distributed (i.i.d.) Gaussian model (Besag, York, and Mollié, 1991). This powerful fusion allows the BYM model to effectively capture both structured spatial variation and unstructured random effects. Embracing INLA unlocks the ability to estimate posterior marginals, skillfully combining prior beliefs, likelihood information, and the observed data to extract profound insights into the intricate landscape of underlying spatial processes.

Structured Random Effect (Spatial Autocorrelation): The structured component captures spatial autocorrelation by incorporating a Gaussian Markov Random Field (GMRF) to model the inherent spatial dependencies within the data. In the BYM context, the structured component is represented as follows:

$$s_i \sim CAR(W) \implies s_i | s_j, j \in N(i) \sim \text{Normal}\left(\frac{\sum_{j \in N(i)} w_{ij} s_j}{\sum_{j \in N(i)} w_{ij}}, \frac{1}{\sum_{j \in N(i)} w_{ij}}\right)$$

Here, s_i represents the spatial random effect for the i-th LSOA region, $N(i)$ is the set of neighboring LSOA regions, and w_{ij} denotes the spatial weight between regions i and j. The prior distribution captures the conditional expectation of s_i considering the spatial random effects of its neighbors, fostering a smooth transition of spatial dependencies.

Unstructured Random Effect (Unexplained Variation): The unstructured random effect accommodates unexplained variability inherent in the data, leveraging a Gaussian distribution. The BYM model entails this unstructured term, which is formulated as:

$$u_i \sim \text{Normal}(0, \tau_u^{-1})$$

In this equation, u_i corresponds to the unstructured random effect for the i-th LSOA region, and τ_u represents the precision parameter governing the distribution.

Priors: Hyperparameters for both structured and unstructured random effects follow a log gamma distribution with shape parameter 1 and rate parameter 5×10^{-4}.

Linear Predictive Equation: Combining structured and unstructured effects within the INLA framework yields the linear predictive equation for the target variable `CURRENT_ENERGY_EFFICIENCY` y_i in region i:

$$y_i = \alpha + x_i^T \beta + s_i + u_i + \epsilon_i$$

Here, α is the intercept, β is the set of regression coefficients associated with the set of covariates x_i, while s_i and u_i denote the structured and unstructured random effects, respectively. The error term ϵ_i signifies the residual noise, drawn from a normal distribution.

Full Model: The complete model, synthesizing both structured and unstructured effects, manifests as:

$$y_i \sim \text{Normal}(\mu_i, \sigma^2)$$

Where μ_i encapsulates the linear combination of covariates, structured, and unstructured effects, and σ^2 signifies the variance parameter.

INLA's Approximation Approach: INLA utilises a combination of analytical approximations and numerical integration to compute approximated posterior marginals of latent Gaussian variable components and hyperparameters within the latent Gaussian model. This approach enables the derivation of comprehensive insights while ensuring computational efficiency.

The integrated ensemble of methods in the INLA BYM model effectively addresses spatial dependencies, unexplained variations, and covariate influences. This results in a sophisticated methodology for predicting continuous target variables within the Bayesian paradigm.

3.4.5 Model Fitting and Prediction

The initial step involves constructing a linear predictor. This predictor formulates the response variable, which represents the current energy efficiency of the properties in question, using a formulation that encompasses the independent variables, referred to as fixed effects.

In the realm of spatial models for lattice data, these models are often defined by incorporating random effects with a variance-covariance structure that depends on the neighborhood structure of the areas. Typically, two areas are considered neighbors if they share boundaries. This adjacency can be as simple as a single point (queen adjacency) or more complex, such as a segment (rook adjacency). In our dataset, the neighborhood structure of LSOA regions was established using the queen adjacency method (refer to Figure 3.8).

Subsequently, the spatial adjacency relationship is transformed into a spatial weights matrix, which captures how each spatial location influences its neighboring areas.

To accommodate both spatial and unstructured effects, a random effect of type 'bym' is introduced. INLA necessitates unique integer values for each distinct data point to model it as a random effect. Consequently, a column labeled 'ID' is generated, corresponding to the row ID of the dataset. The random effect is then defined based on this 'ID' column in conjunction with the spatial weights matrix. The model formula is updated accordingly to incorporate the random effect.

For the rows that require predictions, the response variable values are set to 'NA,' signifying that these areas will not be utilized for model fitting. However, INLA will calculate the predictive distribution, thereby enabling inference on the energy efficiency values in these areas.

Once the model has been fitted, INLA offers options for computing the posterior marginal distributions of the fitted values. This includes calculating the mean or expected value from

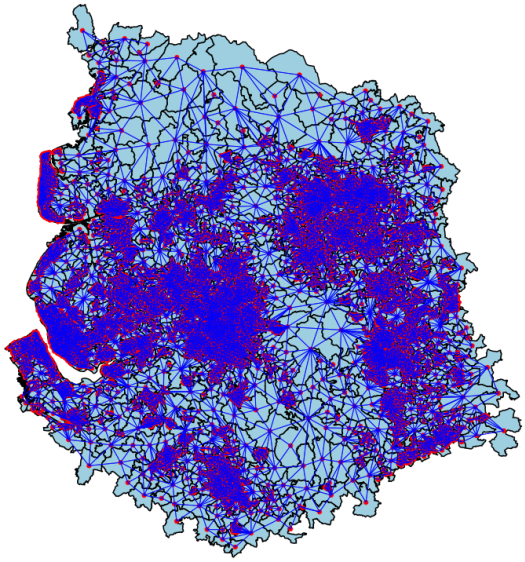

Figure 3.8: LSOA neighbours by queen adjacency

the posterior distribution of each LSOA region, as well as determining the upper and lower limits of the 95% confidence interval. The prediction of energy efficiency for the regions of interest relies on these posterior marginal means and the confidence interval.

Chapter 4

Results

This chapter reveals the culmination of the endeavor, presenting the outcomes of predictions using the INLA framework. It seamlessly integrates Bayesian spatial modeling to forecast energy performance ratings for UK households. Alongside these predictions, a thorough examination of various evaluation metrics sheds light on the models' performance. Navigating through this chapter provides a comprehensive understanding of the model's predictive capabilities, accuracy, and robustness. The content covers the model's feature weights, prediction errors, accuracy measures, and the posterior predictive distribution for energy efficiency. These results unveil the underlying spatial dynamics of energy performance prediction, offering a solid foundation for informed decision-making in the field of energy efficiency modeling.

4.1 Experimental Setup

To evaluate the model's performance, a randomized sampling technique was employed, selecting 30% of the dataset's entries as test regions. The sampling ensured that none of the test regions were immediate neighbors of themselves. The energy efficiency values for these selected entries were deliberately labeled as "NA." This dataset configuration served for both model fitting and prediction generation. Predictions for these unknown values were generated by the model using the associated independent variables while considering the spatial influence exerted by neighboring elements.

In addition, a model featuring only unstructured (i.i.d) random effects was fitted as a baseline. This allowed for assessing the significance of incorporating spatial random effects. Furthermore, an XGBoost model, previously discussed in related works, was trained using the remaining 70% of the data and subsequently evaluated on the designated test regions. This meticulous experimental setup ensures a comprehensive assessment of the model's performance, offering valuable insights into its predictive capabilities and effectiveness.

feature name	**feature importance**
ENERGY_CONSUMPTION_CURRENT	0.584333
MAINHEAT_DESCRIPTION	0.135346
TOTAL_FLOOR_AREA	0.056404
ROOF_DESCRIPTION	0.040814
WALLS_DESCRIPTION	0.036617
FLOOR_DESCRIPTION	0.024701
EXTENSION_COUNT	0.023658
CONSTRUCTION_AGE_BAND	0.020819
PROPERTY_TYPE	0.014463
FLOOR_HEIGHT	0.012041
LIGHTING_DESCRIPTION	0.010751
FLOOR_LEVEL	0.008382
FIXED_LIGHTING_OUTLETS_COUNT	0.007380
TENURE	0.006655
MAINHEAT_CONTROLS	0.005632
WINDOW_DESCRIPTION	0.005445
GLAZED_TYPE	0.004134
HOTWATER_DESCRIPTION	0.002422

Table 4.1: Feature importance table by XGBoost model

4.2 Feature Importance

The analysis of feature importance provides insights into the significance of individual covariates within the predictive models. Different methodologies are employed to gauge feature importance for each model. For the INLA models, the mean value of the posterior marginal distribution of the covariate weights serves as the measure. In contrast, the XGBoost model utilises information gain to quantify feature importance (refer to Table 4.1).

It's noteworthy that in the BYM and IID models, categorical variables are transformed into dummy variables, given the reliance on linear predictors for fixed effects. This results in an extensive set of covariates, including these dummy variables. To maintain brevity, the focus will be exclusively on presenting the top ten features for the INLA models. Additionally, the direction of the weight (positive or negative) is disregarded, with emphasis placed solely on the magnitude. The feature importance tables for INLA models can be found below (refer to Tables 4.2 and 4.3).

feature name	mean	sd	0.025quant	0.5quant	0.975quant
ENERGY_CONSUMPTION_CURRENT	-49.86	0.34	-50.54	-49.86	-49.18
TOTAL_FLOOR_AREA	-15.30	0.33	-15.95	-15.303162	-14.65
FIXED_LIGHTING_OUTLETS_COUNT	13.49	8.07	-2.33	13.49	29.32
MAINHEAT_DESCRIPTIONBoiler and radiators, oil	-5.84	0.96	-7.72	-5.84	-3.95
MAINHEAT_DESCRIPTIONElectric storage heaters	5.22	1.16	2.93	5.22	7.51
FLOOR_HEIGHT	-2.59	0.89	-4.35	-2.59	-0.83
EXTENSION_COUNT	-2.04	0.24	-2.51	-2.04	-1.55
CONSTRUCTION_AGE_BANDEngland and Wales: 2012 onwards	1.69	1.02	-0.31	1.69	3.70
mainheat_controlsmain heat cont group 5	-1.66	0.36	-2.36	-1.66	-0.95
mainheat_controlsmain heat cont group 7	-1.34	0.40	-2.12	-1.34	-0.55

Table 4.2: Important features in BYM model

feature name	mean	sd	0.025quant	0.5quant	0.975quant
FIXED_LIGHTING_OUTLETS_COUNT	61.93	9.05	44.18	61.93	79.67
ENERGY_CONSUMPTION_CURRENT	-45.75	0.36	-46.46	-45.75	-45.04
TOTAL_FLOOR_AREA	-15.77	0.35	-16.46	-15.77	-15.08
FLOOR_HEIGHT	-7.85	0.92	-9.66	-7.85	-6.05
MAINHEAT_DESCRIPTIONBoiler and radiators, oil	-7.59	1.14	-9.83	-7.59	-5.36
MAINHEAT_DESCRIPTIONElectric storage heaters	4.17	1.37	1.49	4.17	6.86
EXTENSION_COUNT	-3.01	0.24	-3.47	-3.01	-2.54
mainheat_controlsmain heat cont group 5	-2.31	0.43	-3.16	-2.31	-1.45
CONSTRUCTION_AGE_BANDEngland and Wales: 2012 onwards	1.86	1.22	-0.54	1.86	4.26
glazed_typetriple glazing	1.38	1.12	-0.82	1.38	3.59

Table 4.3: Important features in IID model

4.3 Prediction Visualisation

The visualisation of predictions for LSOA regions unfolds through the application of three models. Initially, energy efficiencies for the designated LSOA regions of interest undergo prediction using the INLA models and the XGBoost model. An elaborate representation of this prediction process is encapsulated in the figure below (refer to Figure 4.1). Notably, the green-shaded LSOA regions within the first plot correspond to the evaluation-selected test regions.

Subsequently, the INLA models extend beyond point predictions, offering posterior predictive distributions for the LSOA regions. The ensuing figure (refer to Figure 4.2) unveils these distributions, with a focus on the BYM model (left column plots) and the IID model (right column plots). Each row within this visualisation highlights an LSOA region drawn from the test regions. The red lines in each plot represent the actual energy efficiency values for the respective LSOAs. Interestingly, the rows are arranged to align with actual energy ratings, ranging from 'E' to 'B', providing a comprehensive perspective on the predictive capabilities of the models.

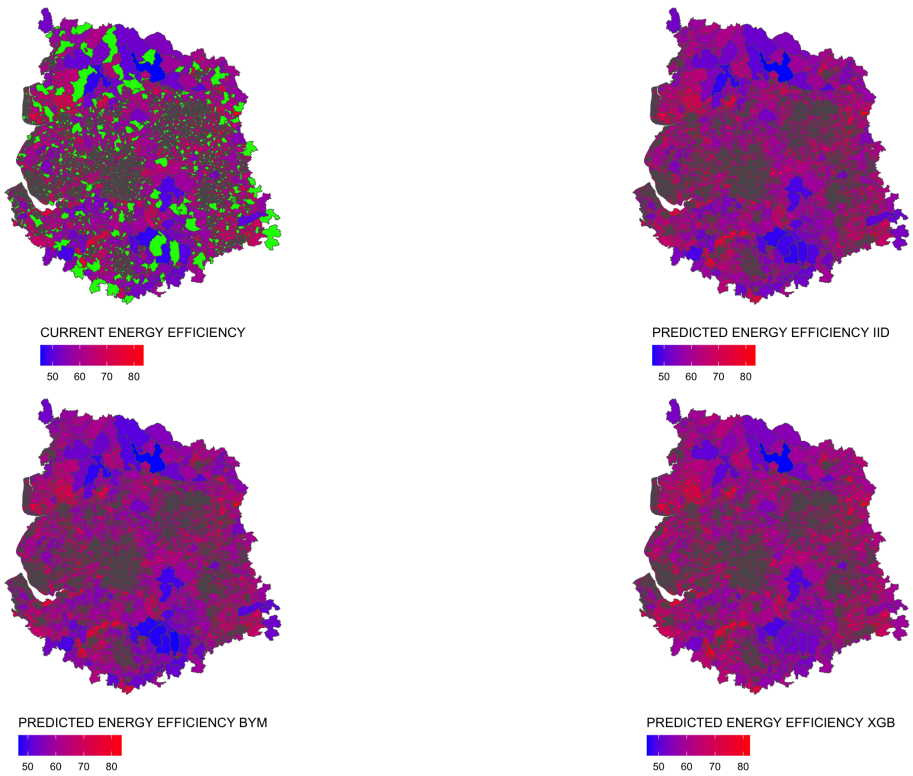

Figure 4.1: Energy Efficiency Predictions over LSOAs

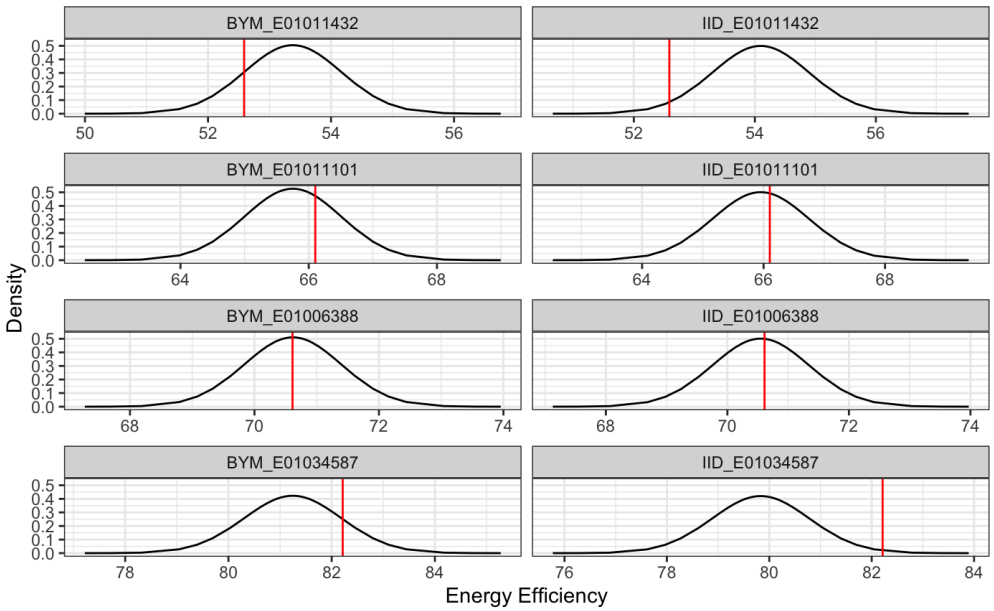

Figure 4.2: Posterior predictive distributions for BYM (left) and IID (right) models

4.4 Prediction Accuracy and Errors

The evaluation of prediction accuracy and errors serves to comprehend model performance. This assessment spans the LSOA regions of interest for the three models utilised in this study: the INLA models incorporating structured (BYM) and unstructured (IID) random effects, and the XGBoost model.

A holistic understanding of model performance is facilitated by visually representing prediction errors. The plot (refer to Figure: 4.3) below offers an overview of prediction errors, calculated as the difference between actual and predicted energy efficiencies, across all LSOA regions. This visualisation illuminates the distribution and variability of prediction errors concerning the geographical layout of the regions.

Quantitative evaluation of model performance is presented via mean squared errors (MSE) computed for each of the three models. These MSE values are derived from the test LSOA regions, providing insight into the average squared deviations between actual and predicted energy efficiencies. The subsequent table (refer to Table: 4.4) summarizes the MSEs of the models.

Additionally, a detailed analysis of prediction accuracy categorises the test LSOA regions into distinct energy rating bands. Energy bands 'B' to 'E' were identified in the test regions. This categorization enables a focused evaluation of each model's performance within specific energy efficiency ranges. The accompanying table (refer to Table 4.5) offers a comparative overview of prediction accuracies across different energy rating bands. Moreover, the table

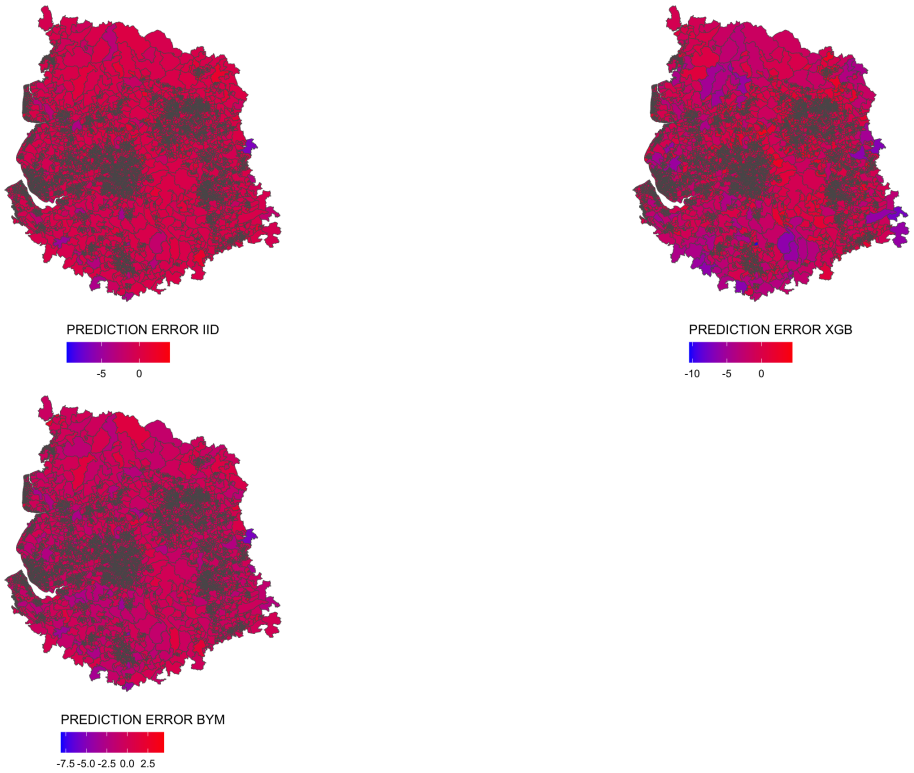

Figure 4.3: Energy Efficiency Prediction Errors over LSOAs

Model	MSE
IID	1.205574
BYM	0.7868315
XGB	1.102648

Table 4.4: Mean Squared Error for test regions

presents the counts of samples within each energy rating band contributing to the testing and training datasets, delivering a comprehensive perspective on dataset distribution and the models' adaptability to diverse energy rating ranges.

actual_band	**accuracy_iid**	**accuracy_bym**	**accuracy_xgb**	**test_samples**	**train_samples**
E	0.7391304	0.7826087	0.7826087	23	72
D	0.9736989	0.9764969	0.9781757	1787	4261
C	0.8782383	0.9015544	0.8963731	386	858
B	0.4000000	0.6000000	0.6000000	5	12

Table 4.5: Band wise accuracy for the models

Through this meticulous analysis of prediction accuracy and error, this section emphasises the strengths and limitations of each model, providing guidance for informed decision-making in the field of energy efficiency modeling and prediction.

Chapter 5

Discussion

This chapter synthesises the discoveries derived from the models' predictions, shedding light on their proficiency in capturing the intricate interplay between covariates and spatial influences. This section engenders a deeper understanding of energy efficiency modeling, charting the course for data-informed decisions and advancements in the field. It achieves this by meticulously balancing empirical evidence with theoretical reasoning.

5.1 Interpreting the effects of features

The significance of various features in predicting energy efficiency becomes a pivotal aspect of our analysis. This section delves into a comprehensive analysis of feature importance, providing insights into the roles played by different variables in determining energy efficiency predictions. The XGBoost model, prioritising predictive accuracy, presents a feature importance table that illustrates the relative importance of various covariates. It's essential to note that a higher absolute value of the feature weight corresponds to a greater degree of importance.

According to the XGBoost model, the pivotal features significantly influencing energy efficiency predictions become apparent. Foremost is the `ENERGY_CONSUMPTION_CURRENT`, which holds a substantial importance of approximately 58%. This underscores the substantial impact of current energy consumption on energy efficiency predictions. Following closely are the `MAINHEAT_DESCRIPTION` and `TOTAL_FLOOR_AREA`, contributing approximately 13% and 5% of importance, respectively, in shaping predictions. In contrast, features such as window description, window glazed type, and hot water description exhibit relatively minor impacts, collectively accounting for the remaining 1%.

While the XGBoost model's feature importance table provides insight into the magnitude of influence, it does not indicate whether these features impact energy efficiency positively or negatively. This is where the INLA model's feature weights come into play. Operating within a Bayesian framework, these weights not only offer magnitude but also direction of

influence on energy efficiency predictions. The resulting posterior distribution of weights instills confidence by illustrating potential ranges in which these weights may fluctuate.

Interestingly, the BYM model aligns with the XGBoost model's findings concerning feature importance. Since the numeric features underwent scaling before fitting, the weights obtained from the model need to be multiplied by the feature range to determine the extent to which a unit change in the feature affects the response. Noteworthy features emerge, unveiling their distinct impacts on energy efficiency predictions. Leading the list is `ENERGY_CONSUMPTION_CURRENT`, displaying a mean weight of -49.8. This corresponds to an estimated 50-unit decrease in the energy efficiency score for every 200 KWh/m^2 increase in annual energy consumption.

`TOTAL_FLOOR_AREA` closely follows, with a mean weight of -15.3. This implies a decrease of approximately 15 units in the efficiency score for every $70m^2$ increase in floor area. Notably,`FIXED_LIGHTING_OUTLETS_COUNT` plays a significant role, with a mean weight of 13.49. This translates to an increase of approximately 13 units in the efficiency score for every 10 efficient lighting outlets added.

The numeric feature `FLOOR_HEIGHT` also emerges, showcasing its effect on energy efficiency. A 3-unit decrease in the efficiency score correlates with every 0.4m increase in height. The type of `MAINHEAT_DESCRIPTION` used in households significantly influences energy efficiency. Notably, boilers and radiators using oil exhibit the least efficiency, resulting in an approximate 6-unit decrease in efficiency score. In contrast, electric storage heaters promote energy efficiency, contributing to an approximate 5-unit increase.

Although categorical features exhibit a lesser overall impact, their insights remain valuable. This analysis not only deepens our understanding of feature importance within the models but also offers practical implications for enhancing energy efficiency predictions and guiding informed decision-making.

5.2 Model Comparisions

An in-depth examination of the three models employed in this study is undertaken by drawing insights from their predictive accuracy and error assessment, as outlined in the results section. The IID model was chosen strategically as a baseline, providing a yardstick to gauge the significance of introducing spatial random effects. A comparison between the Deviance Information Criterion (DIC) and Widely Applicable Information Criterion (WAIC) scores for the BYM model and IID model clearly demonstrates that the inclusion of spatial random effects in the BYM model leads to a notable enhancement in predictive performance.

The evaluation continues with a scrutiny of the mean squared errors (MSEs) for the three models. Remarkably, all three models exhibit MSEs below 2, a commendable feat considering the typical energy efficiency values clustered between 50 and 80. Among them, the BYM model emerges as the frontrunner, boasting the lowest MSE of 0.78. This underscores the model's prowess in capturing and predicting energy efficiency variations.

Actual Band	residuals_iid	residuals_bym	residuals_xgb
E	0.3913043	0.26086957	0.3913043
D	0.1628428	0.09569110	0.1477336
C	0.1088083	0.08549223	0.0880829
B	0.0000000	0.00000000	0.0000000

Table 5.1: Proportion of positive residuals

Actual Band	residuals_iid	residuals_bym	residuals_xgb
E	0.08695652	0.04347826	0.0000000
D	0.16228316	0.10352546	0.1253497
C	0.18652850	0.11917098	0.1735751
B	0.60000000	0.40000000	1.0000000

Table 5.2: Proportion of negative residuals

Further insights are derived from meticulous analysis of prediction accuracies across different energy efficiency bands. Both the BYM and XGBoost models demonstrate consistent accuracy performance throughout the bands. Interestingly, the IID model displays comparable accuracies in bands C and D, but its performance falters in bands B and E. This discrepancy can be attributed to the significantly smaller dataset sizes in bands B and E. The disparity in training data sizes in these bands, with only 12 and 72 data points respectively, underscores the IID model's limitations with limited data.

To ensure rigorous evaluation, residual analysis is employed, considering a one-unit buffer on each side of actual energy efficiencies. The examination of positive and negative residuals across bands for all models highlights a consistent trend: overestimation in lower efficiency bands (see Table 5.1) and underestimation in higher efficiency bands (see Table 5.2). Upon closer examination, it becomes evident that the BYM model exhibits the least over and underestimation compared to the other two models.

In the realm of predictive distributions, the BYM model holds an upper hand. As energy efficiencies are often represented by rating bands, the BYM model's posterior predictive distribution proves to be a favorable choice over the XGBoost model's point prediction. This distinction arises from the BYM model's inherent capacity to consider uncertainties and provide a more comprehensive understanding of the prediction process.

For example, consider the LSOA E01006388, whose actual energy efficiency is 70.6, or a rating of C (ranging from 69 to 80). The BYM model can predict that 95% of the time, the efficiency will fall within the range of 69 to 73, or that the likelihood of energy efficiency being less than 69 or being classified as having a rating of D is only 2%.

5.3 Project Reflection

In conclusion, the comprehensive comparison of predictive models firmly establishes the Besag, York, and Mollié (BYM) model as the most suitable choice for energy efficiency prediction. Throughout the rigorous evaluation of predictive accuracy, error metrics, and residual analysis, the BYM model consistently outperforms other models. The incorporation of spatial random effects within the BYM model not only leads to superior predictive performance, as evidenced by lower Deviance Information Criterion (DIC) and Widely Applicable Information Criterion (WAIC) scores, but also results in the lowest mean squared error (MSE) among all models.

The BYM model's ability to maintain accuracy across diverse energy efficiency bands, coupled with its capacity to account for intricate spatial dynamics through extensive residual analysis, firmly underscores its suitability for energy efficiency forecasting. While the comparison may indicate only a slight advantage over the XGBoost model, the BYM model holds a unique position due to its inherent Bayesian inference capabilities. Given that energy efficiency ratings encompass a range of efficiency scores, Bayesian inference adeptly accommodates uncertainties, offering detailed insights that surpass the point predictions of the XGBoost model.

In this context, the BYM model not only exhibits robustness in capturing intricate spatial variations inherent in energy performance data but also demonstrates its ability to foster a comprehensive understanding. Thus, it emerges as the optimal choice for accurate, insightful, and Bayesian-informed energy efficiency predictions, effectively aligning with the intricate dynamics and uncertainties of real-world energy performance scenarios.

Chapter 6

Limitations and Future Works

In the realm of energy efficiency modeling, it is essential to acknowledge inherent limitations and chart paths toward future refinements. A primary challenge lies in the current spatial modeling approach, primarily centered on Local Super Output Area (LSOA) levels. While effective at capturing general regional trends and spatial relationships, this methodology may fall short in delivering precise predictions for individual households. Consequently, household-level prediction accuracy could be compromised.

This limitation stems from the fact that LSOA-level modeling aggregates properties and their energy efficiency attributes within predefined geographic areas. While this approach captures broad patterns, it may not fully encompass the unique characteristics of individual households. A potential remedy involves adopting more granular spatial units, such as postcode boundaries.

Mitigating this limitation entails the utilisation of the OS Code-Point with Polygons product, facilitating the extraction of postcode boundaries. By incorporating this spatial data, a refined model can be crafted, enabling a deeper understanding of localised trends compared to the broader LSOA methodology. This enhancement holds the promise of significantly elevating the precision of energy efficiency predictions for individual households.

Another avenue for future exploration pertains to the prior distributions employed in the INLA BYM model. Due to project time constraints, a comprehensive exploration of various prior distributions was unfeasible. However, forthcoming research endeavors could involve systematic experimentation with diverse prior distributions tailored to the specific characteristics of energy efficiency data. This investigative approach may yield an even more accurate and dependable predictive model.

By addressing these limitations and pursuing the outlined avenues for future research, the field can advance toward highly accurate, localised energy efficiency modeling. This progression has the potential to unlock new horizons, offering precise predictions and fostering well-informed decision-making in the realm of energy efficiency.

Chapter 7

Conclusions

In conclusion, this research journey has navigated the intricate landscape of energy efficiency modeling, revealing the symbiotic relationship between predictive methodologies and spatial intricacies. Through rigorous analysis and exploration, a comprehensive understanding of predictive models' capabilities, accuracy, and limitations has been unveiled, casting a illuminating light on the complex world of energy efficiency predictions for households in the UK.

By embracing the power of Integrated Nested Laplace Approximation (INLA) models and Gaussian Spatial Processes (GPs), this study has harnessed the essence of Bayesian modeling and spatial dependence. The INLA models, encompassing both intrinsic Conditional Autoregressive (CAR) and i.i.d. random effects, have emerged as robust tools for capturing spatial dependencies. Concurrently, the XGBoost model has showcased its prowess in predictive accuracy and adaptability.

The empirical outcomes have provided remarkable insights. Feature importance analysis has illuminated the pivotal role of diverse covariates, offering a nuanced understanding of their influence on energy efficiency predictions. Through comprehensive comparisons, the strengths of the BYM model and XGBoost have been highlighted, while the limitations of the IID model in smaller datasets have been elucidated.

Moreover, the prediction accuracies and posterior predictive distributions have granted us a profound lens into the predictive capabilities of each model. These insights equip us with a pragmatic understanding of their effectiveness in energy efficiency modeling, allowing for well-informed decisions.

Crucially, the study has underscored the significance of feature importance in identifying energy-efficient choices among categorical features. By dissecting the impact of various covariates, this approach has paved a clear path toward more energy-efficient decisions. This nuanced comprehension empowers stakeholders to strategically navigate the landscape of energy efficiency, capitalizing on the insights gleaned from the models.

Ultimately, this research journey has unfolded against the backdrop of spatial modeling's

intrinsic connection with energy efficiency. It stands as a testament to the potency of spatial intricacies in influencing energy outcomes, and the pivotal role that spatial modeling plays in harnessing these intricacies. As the curtain draws on this journey, spatial modeling stands tall as a beacon of understanding, offering profound insights into energy efficiency dynamics and charting a course toward informed decisions in energy efficiency modeling.

Appendix A

Methodology

A.1 Feature set comprising Energy Performance Certificate

The relevant features that are used in the modelling are given in the table below:

Feature	Description
CURRENT ENERGY EFFICIENCY	energy required for space heating, water heating and lighting [in kWh/year] multiplied by fuel costs
ENERGY CONSUMPTION CURRENT	Current estimated total energy consumption for the property in a 12 month period
TOTAL FLOOR AREA	The total useful floor area is the total of all enclosed spaces measured to the internal face of the external walls
EXTENSION COUNT	The number of extensions added to the property
FLOOR HEIGHT	Average height of the storey in metres
FIXED LIGHTING OUTLETS COUNT	The number of fixed lighting outlets
MAINHEAT DESCRIPTION	Type of main heating system used
CONSTRUCTION AGE BAND	Age band when building part constructed
TENURE	Describes the tenure type of the property. One of: Owner-occupied; Rented (social); Rented (private)
FLOOR DESCRIPTION	Insulation the floor provides
FLOOR LEVEL	Flats and maisonettes only. Floor level relative to the lowest level of the property (0 for ground floor)
GLAZED TYPE	Type of window glazing used
HOT WATER DESCRIPTION	Type of water heating system used
LIGHTING DESCRIPTION	Total number of fixed lighting outlets and total number of low-energy fixed lighting outlets
MAIN HEAT CONTROLS	Type of controller used for the main heating system
PROPERTY TYPE	Describes the type of property such as House, Flat, Maisonette etc
ROOF DESCRIPTION	Insulation the roof provides
WALLS DESCRIPTION	Insulation the walls provide
WINDOW DESCRIPTION	Type and proportion of glazing used for windows

Table A.1: Features in the Energy Performance Certificate that is used in modelling

There are features that are not included for modelling. These include:-

- Feature that represent the potential or estimated values.
- Environmental features
- None of the dates are used as it does not contribute to the energy efficiency

References

Amara, Mohamed and Mohamed Ayadi (Jan. 2012). "The local geographies of welfare in Tunisia: Does neighbourhood matter?*". In: *International Journal of Social Welfare* 22, pp. 90–103. DOI: 10.1111/j.1468-2397.2011.00863.x. (Visited on 07/25/2023).

Banerjee, Anjishnu, David B. Dunson, and Surya T. Tokdar (Dec. 2012). "Efficient Gaussian process regression for large datasets". In: *Biometrika* 100.1, pp. 75–89. ISSN: 0006-3444. DOI: 10.1093/biomet/ass068. eprint: https://academic.oup.com/biomet/article-pdf/100/1/75/481197/ass068.pdf. URL: https://doi.org/10.1093/biomet/ass068.

Besag, Julian, Jeremy York, and Annie Mollié (1991). "Bayesian image restoration, with two applications in spatial statistics". eng. In: *Annals of the Institute of Statistical Mathematics* 43.1, pp. 1–20. ISSN: 0020-3157.

Carr, J. R., R. E. Bailey, and E. D. Deng (Nov. 1985). "Use of indicator variograms for an enhanced spatial analysis". In: *Journal of the International Association for Mathematical Geology* 17, pp. 797–811. DOI: 10.1007/bf01034062. (Visited on 03/20/2020).

Cassie, Noel A C (1993). *Statistics for spatial data : revised edition*. John Wiley & Sons.

Christiaensen, Luc et al. (Nov. 2011). "Small area estimation-based prediction methods to track poverty: validation and applications". In: *The Journal of Economic Inequality* 10, pp. 267–297. DOI: 10.1007/s10888-011-9209-9. (Visited on 04/01/2020).

Communities, The Department of and Local Government (2007). *Energy Performance of Buildings (Certificates and Inspections) (England and Wales) Regulations 2007*. URL: https://www.legislation.gov.uk/uksi/2007/991/memorandum/contents (visited on 09/07/2023).

Energy Performance of Buildings Data England and Wales (2023). epc.opendatacommunities.org. URL: https://epc.opendatacommunities.org (visited on 08/24/2023).

Kass, G. V. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data". In: *Applied Statistics* 29, p. 119. DOI: 10.2307/2986296.

Lindgren, Finn and Håvard Rue (2015). "Bayesian Spatial Modelling with R-INLA". In: *Journal of Statistical Software* 63.19, pp. 1–25. DOI: 10.18637/jss.v063.i19. URL: https://www.jstatsoft.org/index.php/jss/article/view/v063i19.

Lower Layer Super Output Areas (2021) Boundaries EW BFC (2023). geoportal.statistics.gov.uk. URL: https://geoportal.statistics.gov.uk/maps/lower-layer-super-output-areas-2021-boundaries-ew-bfc (visited on 08/24/2023).

Ministry of Housing, Communities and Local Government (2019). *Estimated proportion of the dwelling stock that has had at least one Energy Performance Certificate, financial year ending 2009 to financial year ending 2019, England and Wales*. https://www.ons.gov.uk. URL: https://www.ons.gov.uk/visualisations/dvc765/fig7/index.html (visited on 09/07/2023).

National Statistics, Office for (2021). *Census 2021 geographies - Office for National Statistics*. www.ons.gov.uk. URL: https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeographies/census2021geographies (visited on 07/07/2023).

Open Geography portalx (n.d.). geoportal.statistics.gov.uk. URL: https://geoportal.statistics.gov.uk.

PARLIAMENT, THE EUROPEAN and THE COUNCIL OF THE EUROPEAN UNION (Dec. 2002). *DIRECTIVE 2002/91/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 December 2002 on the energy performance of buildings*. URL: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:001:0065:0071:en:PDF (visited on 09/07/2023).

Postcode to 2021 Census Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (May 2023) Lookup in the UK (2023). geoportal.statistics.gov.uk. URL: https://geoportal.statistics.gov.uk/datasets/postcode-to-2021-census-output-area-to-lower-layer-super-output-area-to-middle-layer-super-output-area-to-local-authority-district-may-2023-lookup-in-the-uk (visited on 08/24/2023).

Rue, Håvard, Finn Lindgren, et al. (n.d.). *R-INLA Project*. www.r-inla.org. URL: https://www.r-inla.org.

Rue, Håvard, Sara Martino, and Nicolas Chopin (Apr. 2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71, pp. 319–392. DOI: 10.1111/j.1467-9868.2008.00700.x. (Visited on 02/09/2022).

Survey, Ordnance (2023). *OS Data Hub*. osdatahub.os.uk. URL: https://osdatahub.os.uk (visited on 08/24/2023).

Williams, Sonia and Christopher Bonham (Feb. 2020). *Using machine learning to predict energy efficiency*. Data Science Campus. URL: https://datasciencecampus.ons.gov.uk/projects/using-machine-learning-to-predict-energy-efficiency/ (visited on 06/20/2023).